AFRL-VS-HA-TR-2004-1111

# LONG-TERM IONOSPHERIC FORECASTING SYSTEM

**Dr. Boris Khattatov**
**Dr. Michael Murphy**
**Dr. Tim Fuller–Rowell**
**Jason Boisvert**

**Environmental Research Technologies**
**1320 Pearl Street, Suite 108**
**Boulder, Colorado 80302**

**30 April 2004**

**Final Report**




**AIR FORCE RESEARCH LABORATORY**
**Space Vehicles Directorate**
**29 Randolph Rd**
**AIR FORCE MATERIEL COMMAND**
**Hanscom AFB, MA 01731-3010**

This technical report has been reviewed and is approved for publication.

| | |
|---|---|
| /Signed/ | /Signed/ |
| JOHN RETTERER | ROBERT A. MORRIS |
| Conract Manager | Branch Chief |

This document has been reviewed by the ESC Public Affairs Office and has been approved for release to the National Technical Information Service.

Qualified requestors may obtain additional copies from the Defense Technical Information Center (DTIC). All others should apply to the National Technical Information Service.

If your address has changed, if you wish to be removed from the mailing list, or if the addressee is no longer employed by your organization, please notify AFRL/VSIM, 29 Randolph Rd., Hanscom AFB, MA 01731-3010. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188



| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| 30-04-2004 | Scientific/Technical Report - Final | 1Jul03 - 31Mar04 |

**4. TITLE AND SUBTITLE**

Long-term Ionospheric Forecasting System

**5a. CONTRACT NUMBER**
F19628-03-C-0077

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Dr. Boris Khattatov

Dr. Michael Murphy, Dr. Tim Fuller-Rowell, Mr. Jason Boisvert

**5d. PROJECT NUMBER**
3005

**5e. TASK NUMBER**
SD

**5f. WORK UNIT NUMBER**
AC

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Environmental Research Technologies, LLC
(a DBA of Fusion Numerics, Inc.)
1320 Pearl Street, Suite 108
Boulder, Colorado 80302

**8. PERFORMING ORGANIZATION REPORT NUMBER**

ERT0010Z

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Air Force Research Lab (AFRL)
29 Randolph Rd.
Hanscom AFB, MA 01731-3010
Contract Manger: John Retterer (VSBXP)

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
AFRL-VS-HA-TR-2004-1111

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for public release; distribution unlimited.

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Report developed under Small Business Innovative Research (SBIR) contract Topic AF03-016. The objective of this effort is to investigate the feasibility of an end-to-end global long-term (up to 3 days) ionospheric forecast model based on a fusion of several diverse technologies and to research the related probability density function (PDF) propagation formalism to characterize the forecast quality. We describe development of the proposed practical system based on a synthesis of several different technologies: (1) an artificial intelligence algorithm known as Support Vector Machines for predicting changes in solar wind from time sequences of solar images; (2) an empirical model of the high-latitude electric field potentials; (3) a physics-based ionospheric model coupled with efficient Kalman filter for forecasting the final ionospheric parameters of interest; and (4) a prototype error propagation scheme based on ensemble filters for computing evolution of forecast probability density functions.

**15. SUBJECT TERMS**
SBIR report, Ionosphere, Modeling, Electron Content, Assimilation, Solar Events, Solar Forecast

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | John Retterer |
| unclassified | unclassified | unclassified | SAR | 31 | 19b. TELEPHONE NUMBER (include area code) (781) 377-3891 |

Unclassified

## Contents

## Figures

## Tables

# 1. INTRODUCTION

The objective of this Phase I effort was to investigate a feasibility of an end-to-end global long-term (up to 3 days) ionospheric forecast model based on a fusion of several diverse technologies and to research the related probability density function (PDF) propagation formalism to characterize the forecast quality.

In order to meet the stated goal of a 3-day forecast one has to address the complete chain of events starting from highly unpredictable and abrupt changes in solar conditions leading to changes in the solar magnetic field, solar wind, etc to solar wind propagation in the interplanetary space to solar wind-magnetosphere interaction to changes in the particle precipitation and electric fields in the ionosphere and finally to changes in neutral and plasma densities.

In the ideal world for each of these processes there would exist a more or less accurate physics-based numerical model. These models would be computationally practical, that is the wall-clock time needed to run these models would be small compared to the time scales of the described phenomena. Additionally, there would exist a sufficient number of spacecraft carrying diverse instruments that supply enough observations to properly constrain and initialize these models with real time conditions. Finally, there would exist a computationally practical method for analyzing the propagation of model and observational errors (PDFs) through each one of these models to the forecasted quantities. The idealized end-to-end system would forecast long-term changes in the ionospheric quantities of interest along with the probabilities (perceived errors) of such forecast.

At the present time such a system is practically impossible. Partly this is due to the fact that some physics-based models (e.g., solar dynamics models) are simply missing or severely deficient. In addition, some of the existing models are computationally impractical. For instance, it might take several days of massively parallel computations running a high-resolution simulation of solar wind propagation from the Sun to 1 AU. By the time such calculations are completed the forecast becomes useless. A fundamental factor hampering both model development and forecasting abilities is the lack of observational data. This factor is the most difficult to overcome as designing and launching the related hardware can take years if not decades and would likely require very significant investment. As progress in numerical weather prediction demonstrated, no forecast can be made without careful analysis of error evolution in the numerical system. The related control space in the case of conventional weather models is already very large (~$10^7$). In the case of space weather the control space will be many more orders of magnitude larger and the error analysis will be either very primitive or computationally prohibitive. A practical formalism for computing PDF evolution in specific space weather applications is missing. Finally, there is a challenge of integrating these complex and computationally expensive models and data streams into a mission-critical real-time forecast system.

There is a significant need for a practical long-term ionospheric forecasting system, yet a proper physics-based assimilative forecasting system, similar in spirit to what is routinely used for numerical weather prediction, will likely not be feasible for years if not decades.

On a high level, our approach to this problem investigated in the Phase I effort is outlined below:

- It is better to do something than nothing; if you don't try you won't get anywhere.
- Don't try to solve everything at once, some pieces will be missing and will have to be filled in later on or by someone else.
- Minimize risk. Try several things and identify candidates that can serve as "building blocks" for a future physics-based end-to-end system.

- When only a marginal forecast can be produced, it is paramount to know its limitations; investigating means of computing error evolution in the system is an important portion of the proposed effort.

- Use physics-based models where such models are sufficiently advanced and/or enough data are available for assimilation. Utilize empirical models and artificial intelligence techniques where physics-based models are missing or are inadequate.

The proposed research involves application of mathematical techniques of estimation theory, inverse problem theory, artificial intelligence and statistics to important and largely unsolved problems in space sciences. We believe that by exploring the interface between these rather distant disciplines we can approach the solution.

In the following sections we will describe technical details of the approach and the core ionospheric model. We then describe the main accomplishments of the Phase I investigation and close with conclusions and recommendations for future work.

The following personnel contributed to this research effort: Dr. Boris Khattatov, PI; Dr. Michael Murphy, Senior Computer Scientist; Mr. Brian Cruickshank, Software Engineer; Jason Boisvert, Research Assistant, and Dr Timothy Fuller-Rowell, Consultant.

## 2. THE APPROACH

The proposed system consists of four major components. These will be discussed in more detail in subsequent sections.

The first component is responsible for predicting environmental conditions at the L1 point given a time sequence of full-resolution images of the Sun. We plan to use time series of SOHO imagery; however we will investigate advantages and disadvantages of using other data sources as well. Examples of SOHO images taken at four different wavelengths are shown in Figure 1. This component is one of the more novel parts of the proposal and will be discussed in more detail than others.

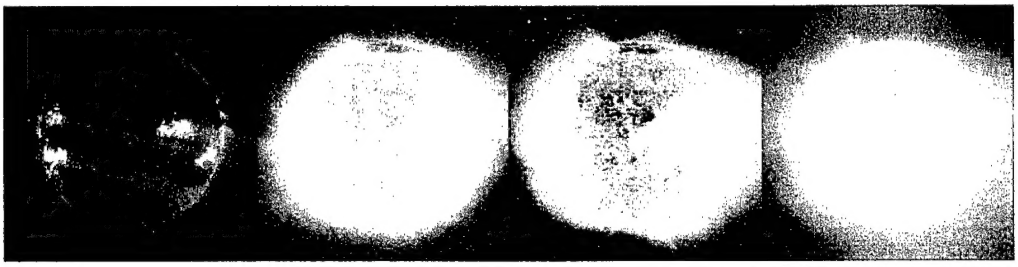


**Figure 1. SOHO Extreme Ultraviolet Imaging Telescope (EIT) images from NASA Goddard Space Flight Center [ 2003/01/10 01:00:15 . Each synoptic full-size image is 1024x1024 pixels.**

We are not going to use physics-based models for this step since at the present time no reliable models of solar dynamics exist and models of solar wind propagation are extremely computationally expensive and still inadequate for predictions. Instead, we will use a machine learning technique known as Support Vector Machines, or SVMs.

SVMs aim to make a prediction of certain quantities based on continuous training given a set of past inputs and outputs. Time series of solar images such as those shown in Figure 1 will serve as inputs and archives of ACE data will constitute the outputs. An example of time series of ACE magnetometer data are shown in Figure 2. Specifics of SVM implementations and creation of the training dataset will be discussed later on.

Obviously, this approach has its limitations. The images have finite resolution (1024x1024 pixels) and finite time discretization, both of which limit the amount of information that they contain. It is not clear whether this information can serve as a predictor of the global coronal magnetic field, which in turn determines future environmental conditions at the L1 point.

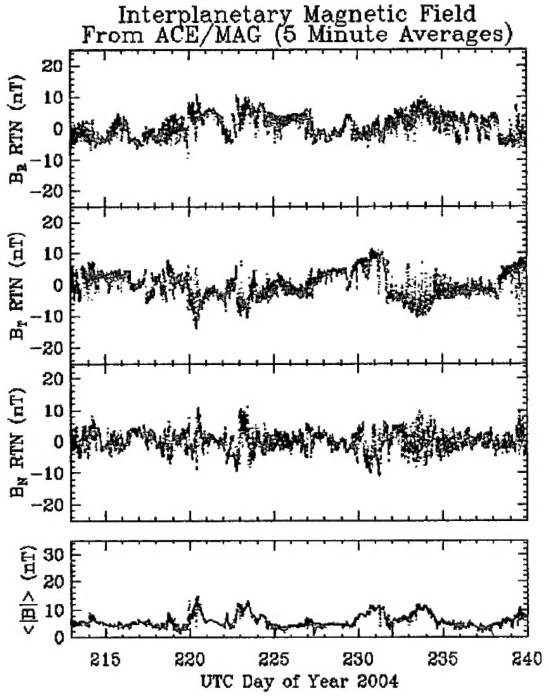



**Figure 2. An example of time series of ACE magnetometer data.**

The coronal magnetic field is unmeasured to this date. Although initiatives are at work today with the goal of measuring it, the first results of these initiatives will be coronal physics, with a long way before products become available that allow the use of this coronal magnetic field for forecasting purposes.

We are also somewhat skeptical of the traditional method which uses the photospheric magnetic field as initial value and extrapolates it up to the corona and beyond:

The usual observations used for this problem have been the "magnetograms." Magnetograms have been proven to be somewhat unreliable approximation to the real field.

The photospheric magnetic field is mapped to a static potential source surface at 2.5 solar radii where the magnetic field is assumed to be radial. Extrapolation techniques are unreliable: tests that compare the extrapolated field 2,000 km above the photosphere with the measured field at such height result in a factor 10 of error. This error can only grow as one requires the magnetic field at bigger distances from the photosphere. The upper corona is so poorly known that the physics which should serve as reference on how to handle the coronal magnetic field and its evolution is a subject of active research.

In the absence of any reliable existing technique to forecast the global coronal magnetic field and, subsequently, the L1 conditions, we believe it is justified to investigate a new method which has not yet been tried for this particular application.

The general idea of the method is that tracers of the magnetic field in the Sun might be used to substitute the lack of knowledge of the solar magnetic field. For instance, sigmoidal structures in the coronal and flares in the chromosphere can give indications of eruptive phenomena with possible effects on space weather. In our case the images are at a 90 degree angle to the observer and the most relevant dynamics happens in the middle of the disc rather than on the perimeter. This makes it harder to identify "tracers" and sigmoidal structures in the image.

Yet, as seen in Figure 1, there clearly is a lot of structure in the middle portion of the solar disk and the proposed machine learning algorithm would attempt to identify which structures likely give rise to enhanced solar wind episodes from observing temporal structure dynamics and the L1 point conditions.

The second component of the system involves predicting ionospheric electric fields from conditions at the L1 point. The approach is outlined in Matsuo et al. (2002). It can be thought of as an extension of the AMIE procedure where Weimer's (2001) empirical model is used to generate climatological high-latitude electric field patterns. EOF decomposition combined with the OI analysis of DE-2 data is used to provide a first order correction.

The third component consists of a physics-based numerical model of the ionosphere coupled with a computationally efficient sub-optimal Kalman filter assimilation scheme that uses real-time TEC data from a network of reference stations. The model has been developed in the course of previous AFRL-funded projects. At present, we are using the system to produce nowcasts of global electron densities via assimilation of GPS reference station data.

The final component of the proposed system is design and implementation of the probability density function evolution. This component should be paramount to any forecast systems as it allows for characterization of the forecast quality and for specifying background error covariances necessary for most data assimilation schemes.

It is safe to say that error propagation schemes specific to most space weather applications simply do not exist. Yet given the present (justified) thrust towards building assimilative models such techniques will have to be developed. Traditionally, error evolution calculations are done using extended Kalman filter methods. We argue that such methods are inappropriate for the space weather applications for two reasons. First, the underlying physical models are mostly undeveloped and will be rapidly evolving. Some perhaps won't exist for decades and empirical methods will be used instead. Extended Kalman filter techniques require building an ad joint model (which has to be repeated every time the model is modified) or running a very large number of finite difference calculations (which is often impractical in space weather). Second, very strong nonlinearities, such as those observed in the solar dynamics and solar wind propagation, make applicability of the extended Kalman filter limited.

Overall, during this project we proposed and explored several fairly high-risk opportunities with the goal of identifying parts that would become building blocks of a future practical forecast system.

We will now describe in more detail the Support Vector Machines and their specific implementation for L1 point observables followed by extended descriptions of the remaining stages.

## 2.1 Support Vector Machines

In 1992, Vapnik introduced the Support Vector Machine (SVM) as an innovative approach to predictor design. Since then SVMs have produced remarkable results on a broad range of important problems in regression, density estimation, anomaly detection, and operator inversion. However, SVMs have made their greatest impact in solving problems of prediction and classification. In particular, relevance to the proposed investigation are the seminal work by Vapnik (Vapnik, 1998) and research on the face recognition problem (Osuna et al., 1997) and chaotic time series forecasting (Mukherjee et al., 1997).

Formally, the classification problem is defined as follows. Let X denote the space of measurements and Y denote the space of class labels. A concept c is a function c: X→Y that determines a label y for each measurement x. A classifier h is a function h: X→Y that approximates the concept. There is no canonical way to generate such a classifier from example data. Indeed, there are many well-known approaches, such as those based on Neural Networks. SVMs offer a novel approach to the classification problem. One salient feature of SVMs is that they cleverly circumvent the curse of dimensionality that plagues other classification methods (including Neural Networks) when working with high-dimensional data sets. Vapnik's innovative approach not only leads to generalization bounds that are independent of dimension but also to a computationally tractable design problem.

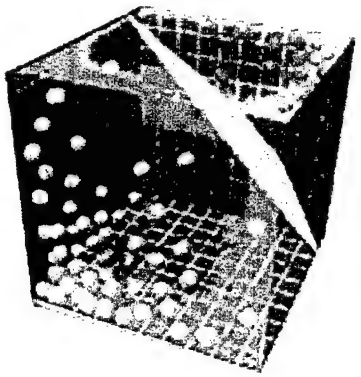


**Figure 3. Support vector machines establish a mapping of input data into a very high-dimensional space where a distinction can be made, for instance, between a set of precursors to a geomagnetic storm and routine solar conditions.**

SVMs combine margin optimization with kernel mappings, which we describe in turn. Consider two concept classes whose data can be linearly separated (i.e., separable by a line in two dimensions, planes in three dimensions, and hyperplanes in higher dimensions). The margin is defined as the distance of the closest sample to the linear separator. Note that the maximal margin classifier is uniquely determined by the closest samples. These samples are called support vectors, and represent the source of the name for this technique. It has long been believed that maximizing the margin improves generalization. Recent theoretical results show that margin not only controls generalization, but does so independently of the dimension of the ambient space. This represents a substantial improvement over existing generalization theory.

It is easy to construct examples of two concept classes for which the data is not linearly separable. However, two concept classes can always be made separable by mapping them to a higher dimensional space. Indeed, SVMs often map to extremely high dimensions in order to obtain a large margin. The motivation for mapping the data into a high-dimensional feature space is that linear decision boundaries constructed in the high-dimensional feature space correspond to nonlinear decision boundaries in the input space.

Given a training set of $M$ samples $\{x_1, x_2, \ldots, x_M\}$ with known class labels $\{y_1, y_2, \ldots, y_M\}$, a new data point is assigned a label by SVM according to the decision function:

$$f(x) = sign\left(\sum_{i=1}^{M_S} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b\right)$$

where

$$k(\mathbf{x}_i, \mathbf{x}) = \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \right\rangle$$

is the kernel function that defines the feature space, $\Phi(\mathbf{x})$ is a nonlinear mapping function from input space to feature space, $b$ is a bias value, and $\alpha_i$ are positive real numbers obtained by solving a quadratic programming (QP) problem that yields the maximal margin hyperplane (Vapnik, 1998). A potential disadvantage of such a mapping is that calculations in this space may be computationally prohibitive. However, the only operation required by SVMs in the mapped space is the inner product. Mappings for the inner product between two points in the image can be evaluated using a bivariate function on the original space, which are known as kernel mappings. SVMs use kernel mappings to implement large margin linear classifiers in extremely high dimensions while performing computations in the feature space.

### 2.1.1 Solar feature-extraction methods

How does one apply SVMs to prediction of geomagnetic disturbances? Because SVMs, as well as other machine-learning algorithms, use numerical values as inputs, the first problem is to codify a sequence of images of the Sun and turn them into vectors.

Each image pixel is treated as a component of a very large dimensional vector of dimension equal to the number of pixels in the image times the number of spectral components times the number of images in the training time sequence of images. Note that as an additional re-processing measure Images might need to be de-rotated to remove the effects of the solar rotation.

Next, this vector will be archived together with the corresponding scalar quantities ($B_x$, $B_y$, $B_z$, solar wind velocity, ...) from the ACE satellite measurements and relevant auxiliary information. Initially, the ACE data will be extracted while taking into account the time delay due to the finite propagation speed of the solar wind. Later on the time delay can be made a parameter in the classifier scheme. We archived all quantities available from ACE. However, only a subset of these might actually be predictable given the classifier inputs.

Once we obtained the vector corresponding to the image sequence, we need to reduce its dimensionality in order to provide a practically feasible input for SVM classifier training.

Originally we proposed using a principal component analysis (PCA) by performing eigenvector decomposition of the matrix obtained via multiplying this vector by its transpose. It is well known in the studies of face recognition that the first several eigenvectors are usually the same for all patterns belonging to the same type (e.g., reflecting the fact that the Sun is round and the image itself is square). These eigenvectors do not contain any new information and can be discarded. The remaining set of eigenvectors can be re-arranged in a vector of size smaller than the original since by now we removed all irrelevant elements. This vector can serve as the input to the SVM classifier for a given sequence of images. We attempted this approach in the course of this project and learned that it is not suitable for dimensionality reduction. The set of different patterns on the solar disk appears to be so diverse that it simply cannot be represented by a linear combination of a relatively small number of "basis functions." We therefore resorted to a different method of dimensionality reduction described later in this document.

When implementing the classifier scheme we attempted to answer only one question: did a particular

time sequence of solar images result in significantly different conditions at the L1 point later on?

The definition of what constitutes a significantly different condition can clearly be made different; the particular definition adopted here (and described in detail later) referred to a sudden increase in either bulk solar wind speed or total magnitude of the interplanetary magnetic filed (IMF) at the L1 point one to 4 days after the particular change in solar images took place.

### 2.1.2 Design of a kernel function

The other question to consider is the kernel function arising in the SVM formalism. For this, there is a large set of choices. Indeed the ability to tailor the kernel mapping to the problem at hand is another prominent feature of SVMs. Many different kernel functions are discussed in (Osuna et al., 1997). It is reported that SVM classification model based on the linear kernel often performs well for feature selection and one might try to avoid nonlinear kernels in order to maintain model simplicity. We will therefore start our simulations using the linear kernel. Linear kernel is a special case of the polynomal kernel given by:

$$k_P(\mathbf{x}_i, \mathbf{x}_j) = (a + b\mathbf{x}_i \bullet \mathbf{x}_j)^d$$

when $a = 0$, and $b = d = 1$.

The problem of selective kernel scaling is central to the success of the SVM classification model. Selective scaling makes it possible to assign a different scaling factor to each input feature variable based on its importance to the classification problem.

## 2.2 Ionospheric Electric Fields

The predictability of the driven thermosphere-ionosphere system is inherently dependent on an accurate knowledge of its forcing.

In order to specify the high-latitude ionospheric electric field at a relatively small computational cost, we considered the approach based on the optimal interpolation method of the Assimilative Mapping of Ionospheric Electrodynamics (AMIE). The approach is similar to the AMIE procedure developed by Richmond et al., (1988) but without any assumption on the ionospheric conductivity distribution.

In the proposed approach a set of 11 Empirical Orthogonal Functions (EOFs) derived from in-situ measurements of plasma drifts from the Dynamic Explorer 2 (DE-2) satellite by Matsuo et al. (2002) is used as a basis functions.

The use of EOFs as bases enables one to incorporate realistic spatial coherence of the electric field on a large scale into the analysis so that the data void area will be well constrained; furthermore, it reduces the background error covariance to a diagonal matrix and the required number of basis functions. Now, the state variable is the coefficients of the EOF bases and the observation operator is expressed in terms of the EOFs as well.

In order to specify the prior (forecast) distribution of the electric field, the climatological electric potential model of Weimer (2001) can be invoked with appropriately time-delayed solar wind conditions estimated/measured at the L1 point.

This time-delay takes into account the solar wind travel time to the magnetopause location from L1 point as well as the ionospheric convection response time to IMF variations.

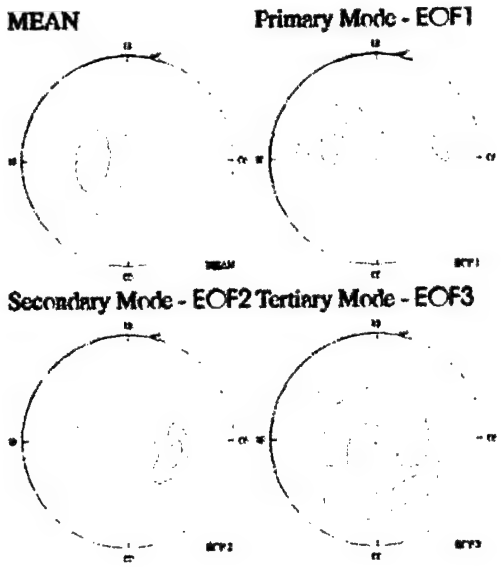



**Figure 4. An example of the mean and the first 3 EOFs calculated by the procedure described in Matsuo et al., (2002).**

In order to fully incorporate the time-dependence, one requires a forecast model that projects the current state (described in terms of the EOFs) to the future state. The property of this forecast model will be obtained from a nonparametric regressional analysis of coefficients of the EOFs from nowcasting and solar wind conditions at the L1 point. Once the forecast model is derived, the OI method will be extended to sequential methods so that the background error covariance evolves explicitly and sequentially at each analysis time step, reflecting preceding analyses that contain all information of observations antecedent to the current analysis.

While this component is clearly necessary for a practical end-to-end forecasting system we had to postpone its development to the Phase II project for two reasons not related to the scientific feasibility of this approach:

Dr. Tomoko Matsuo, who was supposed to be leading development of this component, accepted a postdoctoral position elsewhere. While we made arrangements for Dr Matsuo to participate in the development on a part-time basis at the beginning of the investigation, Dr Matsuo later decided to cancel her involvement.

The DMSP data that we planned to use for the development of this model was made available to us only 3 months before the scheduled project end date.

Given these two complications we decided to devote more resources to the development of the SVM classifier, which looked more promising. We also made arrangements with Dr Barbara Emery of NCAR High Altitude Observatory to include the DMSP data to the CEDAR community database and to assist us with development of this component during the Phase II effort.

## 2.3 The Model

The developed model is a numerical global model of the ionosphere system loosely based on Millward et al (1996), Bailey and Balan (1996), Fuller-Rowell (1996) and Huba et al (2000).

The dynamic equations and vertical ExB transport for seven ions (H, O, $O_2$, He, NO, $N_2$, N) are solved on a fixed Eulerian grid in magnetic p, q, and longitude coordinates. An example of the low-latitude

configuration is shown below (only 20 longitudes and 30 p values are shown for clarity, regular model configuration is 100x100x100).

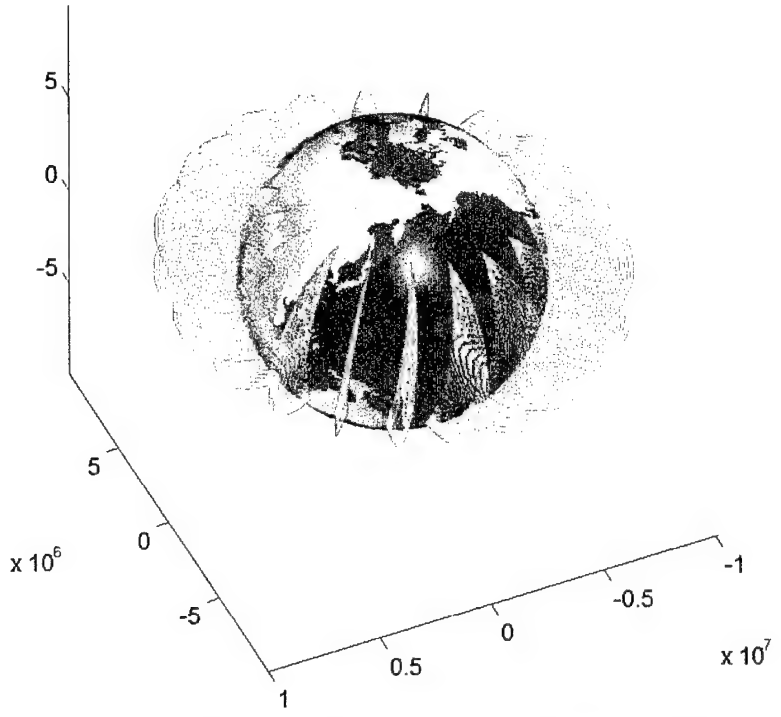



**Figure 5. A Part of the Model Magnetic Grid.**

The model solves plasma dynamics equations – parallel and ExB continuity and momentum – for seven ion species and electrons and energy conservation equation for the three major ions and electrons. The model includes chemical interactions with neutrals and ion-ion and ion-neutral collision rates and Photoionization. ExB drift is computed by the Fejer&Scherless model at the equator and by Weimer (2000) model at high latitude. The high-latitude transport is currently turned-off pending model validation at low latitudes.

### 2.3.1 Prognostic Equations

A prognostic equation allows one to estimate a particular prognostic variable at a future time. The prognostic variables are density, velocity and temperature for ions and electron density, temperature and velocity.

These equations are given in dipole coordinates, along magnetic flow tubes. Therefore, there is only one dependent spatial coordinate corresponding to the position along the magnetic flow tube. This can be a non-dimensional variable q or a dimensional variable $s = q \cdot \mathrm{R_e}$ (Re is the radius of the earth).

Model prognostic, external and diagnostic variables are listed in Tables 1-3.

## Table 1. Model Prognostic Variables

| Variable Name | Units | Description | Comments |
|---|---|---|---|
| Ion densities | particles/ $m^3$ | Local (point) volume density of a particular ion species | at present there are 7 ions: $O^+$, $H^+$, $He^+$, $N_2^+$, $O_2^+$, $NO^+$, $N^+$ |
| Ion temperatures | K, degree | Local temperature of a particular ion species | same |
| Ion velocities | m/s | Local (point) velocity of a particular ion species along the magnetic field line passing through this point | same |
| Electron temperature | K, degree | Local electron temperature | |
| Electron velocity | m/s | Local (point) velocity of electrons along the magnetic field line passing through this point | |
| Electron density | particles/ $m^3$ | Local (point) volume density of electrons | is the sum of all local ion densities |

## Table 2. Model External Variables

| Variable Name | Units | Description | Comments |
|---|---|---|---|
| Neutral densities | particles/ $m^3$ | Local (point) volume density of a particular neutral species | - at present there are 7 neutrals: O, $O_2$, $N_2$, He, H, NO, N. |
| Neutral temperature | K, degree | Local temperature of all neutral species (one for all) | same |
| Neutral zonal velocity | m/s | Local (point) velocity of all neutral species (one for all) in the zonal direction (east-west, eastward is positive) | same |
| Neutral meridional velocity | m/s | Local (point) velocity of all neutral species (one for all) in the meridional direction (north-south, northward is positive) | same |

Table 3. Model Diagnostic Variables

| Variable Name | Units | Description | Comments |
|---|---|---|---|
| ExB zonal velocity | m/s | Local (point) velocity associated with the zonal ExB drift of the magnetic field line passing through this point | Needs to be computed from empirical ExB models |
| ExB meridional velocity | m/s | Local (point) velocity associated with the meridional ExB drift of the magnetic field line passing through this point | same as above |
| Photo production | Particles/s/$m^3$ | Number of particles of a particular ion species produced as a result of photoionization per second per unit volume. | - at present there are 7 neutrals, only 5 of those can be photoionized: O, $O_2$, $N_2$, He, N. Other ions are produced via chemical reactions, such as $O^+ + H \rightarrow H^+ + O$. |
| Chemical production | Particles/s/$m^3$ | Number of particles of a particular ion species produced as a result of chemical reactions per second per unit volume. | There are 21 chemical reactions at the present. e.g., $O^+ + H \rightarrow H^+ + O$. |
| Chemical loss | Particles/s/$m^3$ | Number of particles of a particular ion species lost as a result of chemical and recombination reactions per second per unit volume | This value is a product of the density (concentration) of the ion species being destructed and the *chemical loss rate*, $\boldsymbol{L}$. |
| Photoionization rates | 1/s | Coefficients needed to compute photo production | - at present there are 7 neutrals, only 5 of those can be photo-ionized: O, $O_2$, $N_2$, He, N. |
| Chemical reaction rates | $m^3/s$ | Coefficients needed to compute chemical loss due to electron exchange reactions. | There are 21 chemical reactions at the present. e.g., $O^+ + H \rightarrow H^+ + O$. |
| Recombination reaction rates | 1/s | Coefficients needed to compute loss due to recombination chemical reactions | There are 7 recombination reactions, e.g., $O^+ + e \rightarrow O$. e represents an electron. |
| Ion-neutral collision frequencies | 1/s | Drag on a particular ion particle due to collisions with a neutral species. | There are 7 ions and 7 neutrals, therefore it is a 7x7 matrix with zero diagonal. |
| Ion-ion collision frequencies | 1/s | Drag on a particular ion particle due to collisions with a different ion species. | There are 7 ions, therefore it is a 7x7 matrix with zero diagonal. |
| Ion heating rates | $J/m^3/s$ | Heating due to Joule heating, frictional collisions and other processes. | Is only computed for three major ions, $O^+$, $H^+$, $He^+$ |
| Ion thermal conductivities | J/K/m/s | | Is only computed for three major ions, $O^+$, $H^+$, $He^+$ |
| Electron heating rates | $J/m^3/s$ | Heating due to Joule heating, frictional collisions and other processes. | |
| Electron thermal conductivities | J/K/m/s | | |

### 2.3.2 Continuity Equation For Each Ion Species

Numerical solution of this equation should generate ion density $N_i(t+\Delta t)$ given all related variables at time *t*.

$$\frac{\partial N_i}{\partial t} - b_s^2 \frac{\partial\left(\frac{N_i V_i}{b_s}\right)}{\partial s} + N_i \cdot \nabla V_\perp + \nabla N_i \cdot V_\perp = P_i - L_i \cdot N_i \qquad (1)$$

where

$N_i$ – density of ion i

$V_i$ – velocity (aligned with the magnetic flow tube) of ion i

$s = q \cdot \mathrm{R_e}$

$b_s = \sqrt{1+3\cos^2(eccLat)} \cdot \left(\frac{\mathrm{R_e}}{eccRadius}\right)^3$

$P_i$ – chemical production + photochemical production

$L_i$ – chemical loss rate

$L_i \cdot N_i$ – chemical loss

The term

$$\nabla V_\perp = \frac{6 \cdot V_\perp^{eq} \sin^2(eccLat) \cdot (1+\cos^2(eccLat))}{p \cdot R_e \cdot (1+3 \cdot \cos^2(eccLat))^2} \qquad (2)$$

is a divergence of ExB velocity in the vertical (and meridional) plane, i.e., in p direction. $V_\perp^{eq}$ is the value of ExB meridional drift at the magnetic equator corresponding to a particular p.

### 2.3.3 Momentum Equation For Each Ion Species

Numerical solution of this equation should generate ion velocity $V_i(t+\Delta t)$ given all related variables at time *t*.

$$V_i = \frac{1}{\sum_{n=1}^{N\_Neutrals} \nu_{in} + \sum_{j=1}^{N\_Ions} \nu_{ij}} \cdot \left[ \begin{array}{l} -g \sin I + \frac{b_s k_i}{m_i}\left(\frac{T_i}{N_i}\frac{\partial N_i}{\partial s} + \frac{T_e}{N_e}\frac{\partial N_e}{\partial s} + \frac{\partial(T_i+T_e)}{\partial s}\right) \\ + \sum_{n=1}^{N\_Neutrals} \nu_{in}(V_n \cos D - U_n \sin D)\cos I + \sum_{j=1}^{N\_Ions} \nu_{ij} V_j \end{array} \right] \qquad (3)$$

where

$N_i$ – density of ion i

$V_i$ – velocity (aligned with the magnetic flow tube) of ion i

$V_j$ – velocity of ion j.

$$s = q \cdot \mathrm{R_e}$$

$$b_s = \sqrt{1+3\cos^2(eccLat)} \cdot \left( \frac{\mathrm{R_e}}{eccRadius} \right)^3$$

$m_i$ – mass of ion i.

k – Boltzmann's constant.

$T_i$ – temperature of ion i.

$T_e$ – electron temperature

$N_e$ – electron density.

$U_n$ – zonal neutral velocity.

$V_n$ – meridional neutral velocity.

$\nu_{in}$ - ion-neutral collision frequency.

$\nu_{ij}$ - ion-ion collision frequency.

g – acceleration of gravity.

I – inclination angle for this flow tube:

$$\sin I = \frac{2\cos(eccLat)}{\sqrt{1+3\cos^2(eccLat)}}$$

$$\cos I = \frac{\sin(eccLat)}{\sqrt{1+3\cos^2(eccLat)}}$$

D – declination angle for this flow tube.

### 2.3.4 Energy Equation For Each Ion Species

Numerical solution of this equation should generate ion temperature $T_i(t+\Delta t)$ given all related variables at time *t*.

$$\frac{3}{2} k N_i \left( \frac{\partial T_i}{\partial t} + V_\perp \nabla T_i \right) = k N_i T_i b_s^2 \frac{\partial}{\partial s}\left( \frac{V_i}{b_s} \right) - k N_i T_i \cdot \nabla V_\perp + b_s^2 \frac{\partial}{\partial s}\left( \kappa \frac{\partial T_i}{\partial s} \right) + \frac{3}{2} k N_i V_i b_s \frac{\partial T_i}{\partial s} + Q + F \quad (4)$$

Where

Q

and F - are the heating rates

$\kappa$ -

is the thermal conductivity

### 2.3.5 Electron Temperature Equation

It is similar to ion temperature equation, except that the conductivities and heating rates are computed for electrons.

### 2.3.6 Electron Density Equation

$$N_e = \sum_{i=1}^{NumberOfIons} N_i \qquad (5)$$

### 2.3.7 Electron Velocity Equation

Assumes that there are no field-aligned currents:

$$V_e = \frac{\sum_{i=1}^{NumberOfIons} V_i N_i}{N_e} \qquad (6)$$

Some of the model results are shown in the figures below.

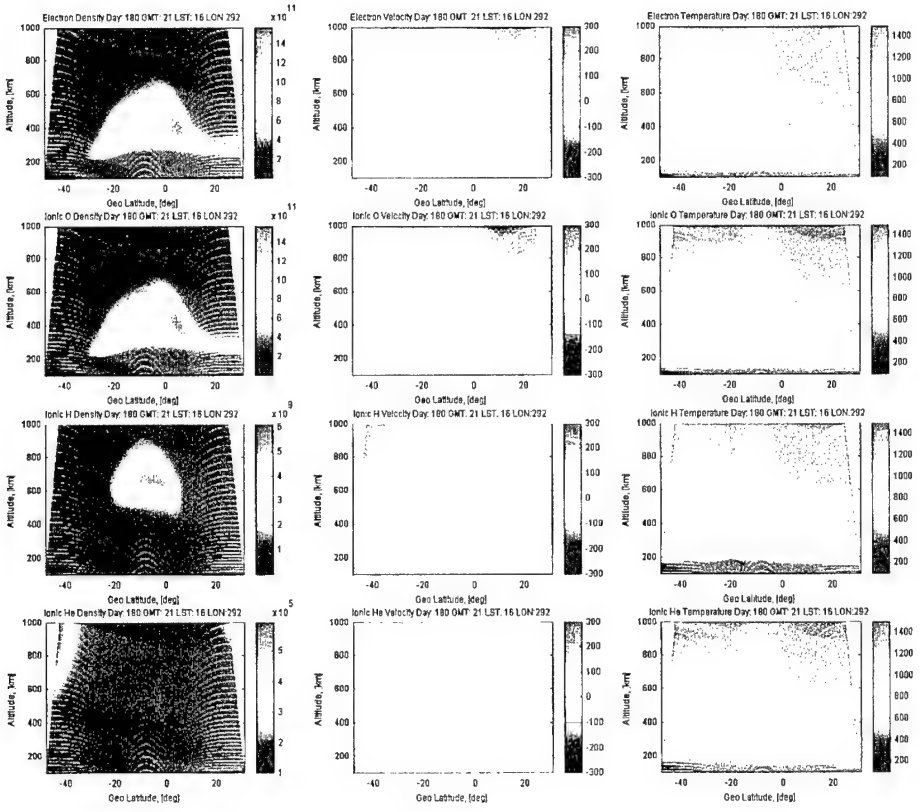



**Figure 6. Examples of instantaneous model prognostic variables for electrons and major ions shown as q-p cross sections at a fixed magnetic longitude.**

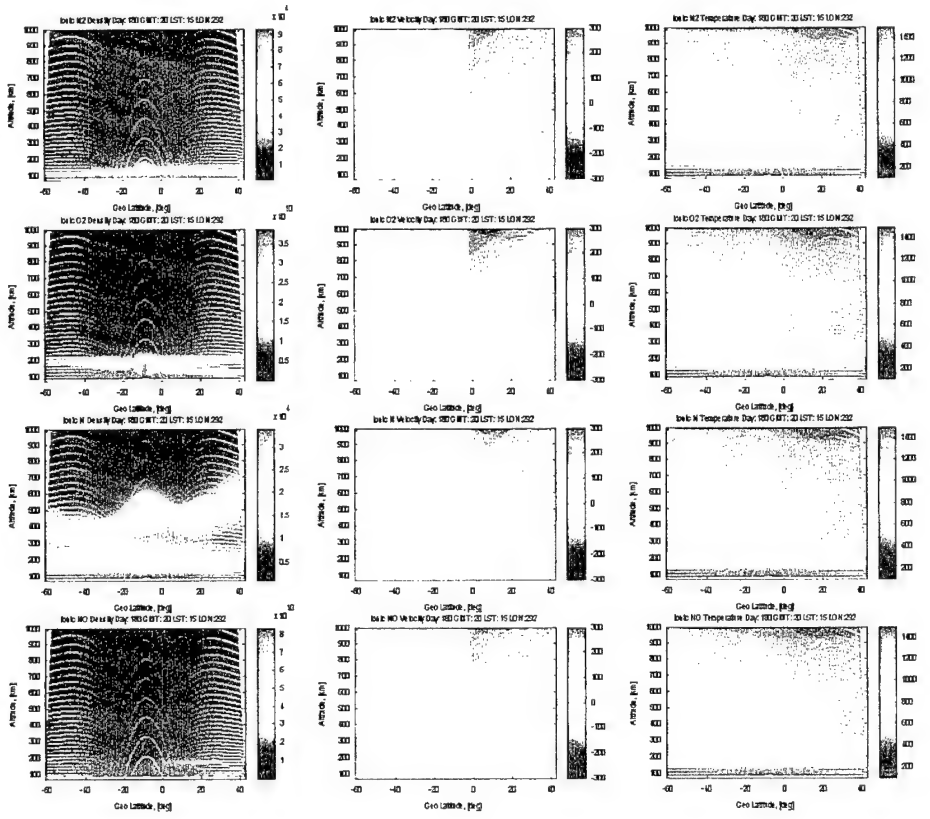


**Figure 7. Examples of instantaneous model prognostic variables for minor ions shown as q-p cross sections at a fixed magnetic longitude.**

## 2.4 PDF Evolution and Ensemble Filters

Kalman filtering is a powerful framework for solving data assimilation problems. By using a Kalman filter, the information provided by the resulting stochastic–dynamic model and the noisy measurements are combined to obtain an optimal estimate of the state of the system.

The ensemble Kalman filter (EnKF) was introduced by Evensen (1994) and has been used successfully in many applications (see Evensen and Van Leeuwen 1996; Houtekamer and Mitchel 1998; Canizares 1999). This Monte Carlo approach is based on a representation of the probability density of the state estimate by a finite number *N* of randomly generated system states. The algorithm does not require a tangent linear model and is very easy to implement. The computational effort required for the EnKF is approximately *N* times as much as the effort required for the underlying model. Figure 8 illustrates the ensemble Kalman filter method.

In a nutshell, the model control space is perturbed according to its perceived uncertainties and an ensemble of possible states is created. Then a set of model integrations is performed and the PDF of the final state is estimated. If data are available the final state and the final PDFs can be corrected using linear regression and the observational operator H and the cycle is repeated. The main challenge is to select a sufficiently small number of ensemble members to make computations practical and yet properly sample the possible control space.

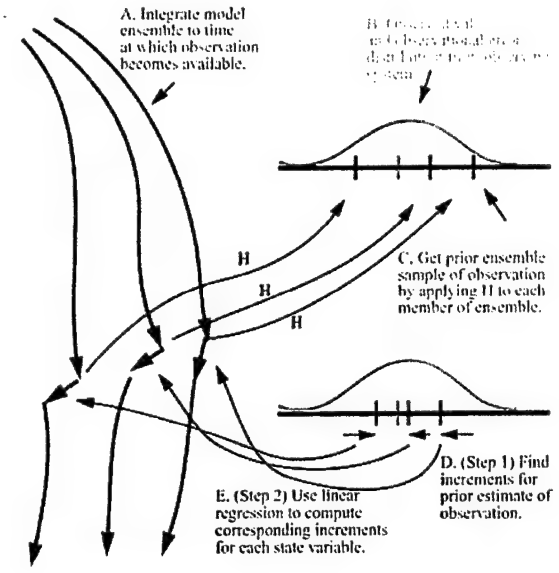



**Figure 8. How ensemble Kalman Filters work (after J. Anderson).**

The only serious disadvantage is that the statistical error in the estimates of the mean and covariance matrix from a sample decreases very slowly for a larger sample size. This is a well-known fundamental problem with all Monte Carlo methods. As a result, for most practical problems the sample size chosen has to be rather large. Here it should be noted that a properly constructed ensemble Kalman filter can still provide an improved analysis even with small-sized ensembles (see Houtekamer and Mitchel 1998). Another approach to solve large-scale Kalman filtering problems is to approximate the full covariance matrix of the state estimate by a matrix with reduced rank. This approach was introduced by Cohn and Todling (1995, 1996) and Verlaan and Heemink (1995, 1997), where the latter used a robust square root formulation for the filter implementation. Algorithms based on similar ideas have been proposed and applied by Lermusiaux (1997) and Pham et al. (1998).

The reduced-rank approaches can also be formulated as an ensemble Kalman filter where the $q$ ensemble members have not been chosen randomly, but in the directions of the $q$ leading eigenvectors of the covariance matrix (see Verlaan and Heemink 1997). As a result these algorithms do not require a tangent linear model. The computational effort required is approximately $q$+1 model simulations plus the computations required for the singular value decomposition to determine the leading eigenvectors [O($q$3); see Heemink et al. (1997)]. In many practical problems the full covariance can be approximated accurately by a reduced-rank matrix with relatively small value of $q$.

Some advantages of Ensemble Filters over Extended Kalman filters are:

Basic update algorithms require only 10's of lines of code.

Require no auxiliary information about model like an adjoint.

Do not rely on linear approximation.

Produce information about complete probability distribution.

Efficient use of observation information with time-varying covariances.

A serious advantage of such an approach is that the filter needs only be implemented once and the same PDF evolution and data assimilation scheme should work with all three components. If one or several components change, say, as a result of a new scientific development, the filter component should still work as before.

Moreover, we argue that in a situation where a mix of machine learning, empirical, and first-principles based models is present, ensemble filters are perhaps the only possible means of computing PDF evolution as building an adjoint of SVM classifier clearly does not make much sense. Finally, in the ensemble filter it is very easy to augment the control space with new parameters, thus incorporating contributions from errors in initial conditions, unknown inputs and model and observational errors.

# 3. RESULTS

## 3.1 Support Vector Machine Classification

At this stage in our research, several methods of organizing the EIT data to train SVM hold considerable promise. The most successful method so far has been to encode the changes between consecutive images as the differences between averaged blocks of pixels.

In essence, we gave the SVM a resolution-reduced view of the differences between consecutive solar images. Each image was divided into 4096 blocks of 8x8 pixels, which were averaged and placed as stacked rows into a single column vector. The difference between such vectors for several consecutive images was taken as a minimal data unit for SVM.

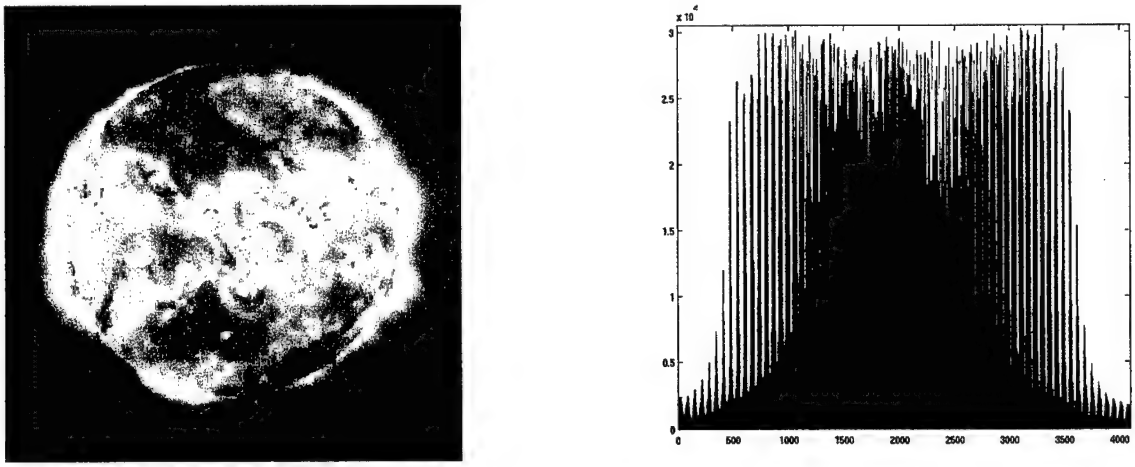



**Figure 9. EIT image and corresponding averaged blocks from Oct., 10, 2002 at 16:48.10 UTC.**

A block size of 8x8 pixels was chosen because changes in the images on regions smaller than this showed essentially no resulting changes in the ACE data. In this sense, the size of the block was our first filter of small-scale changes. Moreover, since EIT images are plagued with missing data, by averaging we could quickly recognize (and ignore) blocks where pixels were missing. Missing pixel blocks are given values of zero in the calibration process, so if either of the corresponding averaged

blocks in consecutive images were equal to zero, we assumed data was missing in both images and set the corresponding difference to zero.

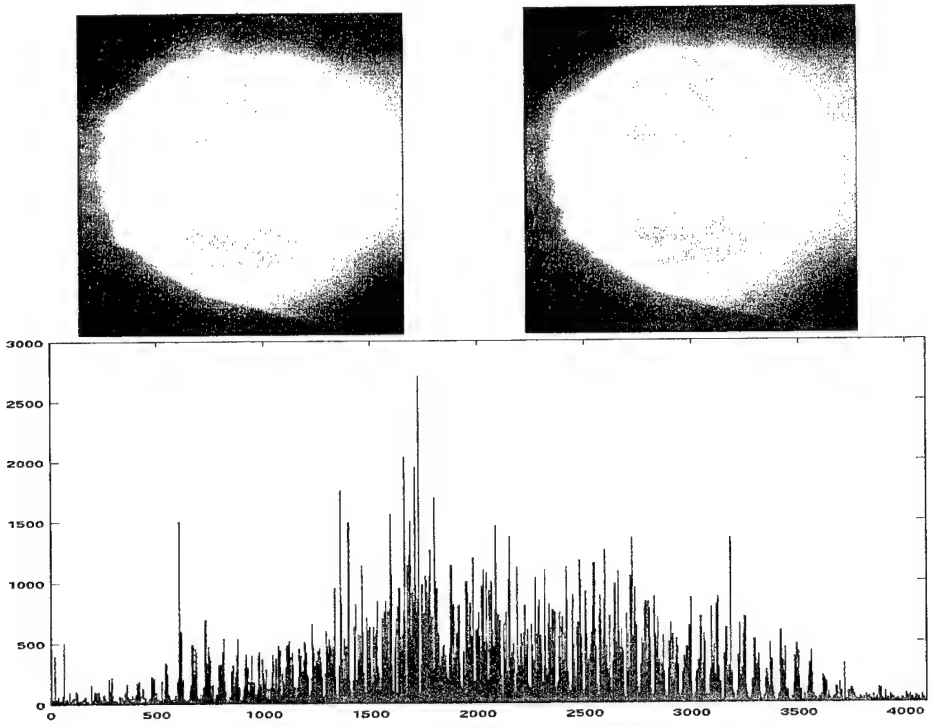



**Figure 10. EIT images and corresponding difference vector from March, 21, 2002 at 08:48.10 09:00.10 UTC. This is considered to be a non-event situation.**

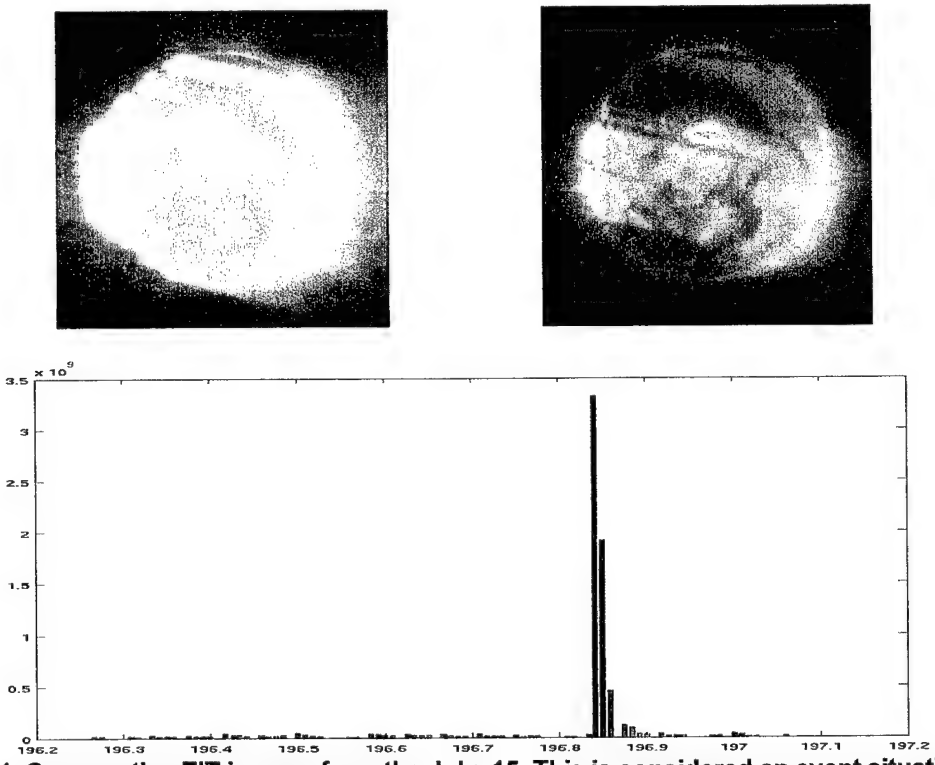



**Figure 11. Consecutive EIT images from the July, 15. This is considered an event situation.**

The images are always changing to some degree, and the regions we're most interested in tracking are always considerably greater in intensity than their surroundings. Taking advantage of SVM's efficient use of long, sparse vectors, we could establish a lower bound on the amount of change within a block necessary to be associated with an event. Therefore, if the difference between matching blocks in consecutive images was not above the level of background changes, we set the value at that index to zero. A vector with 4096 elements could be reduced at times to a set of a few values and their corresponding indices.

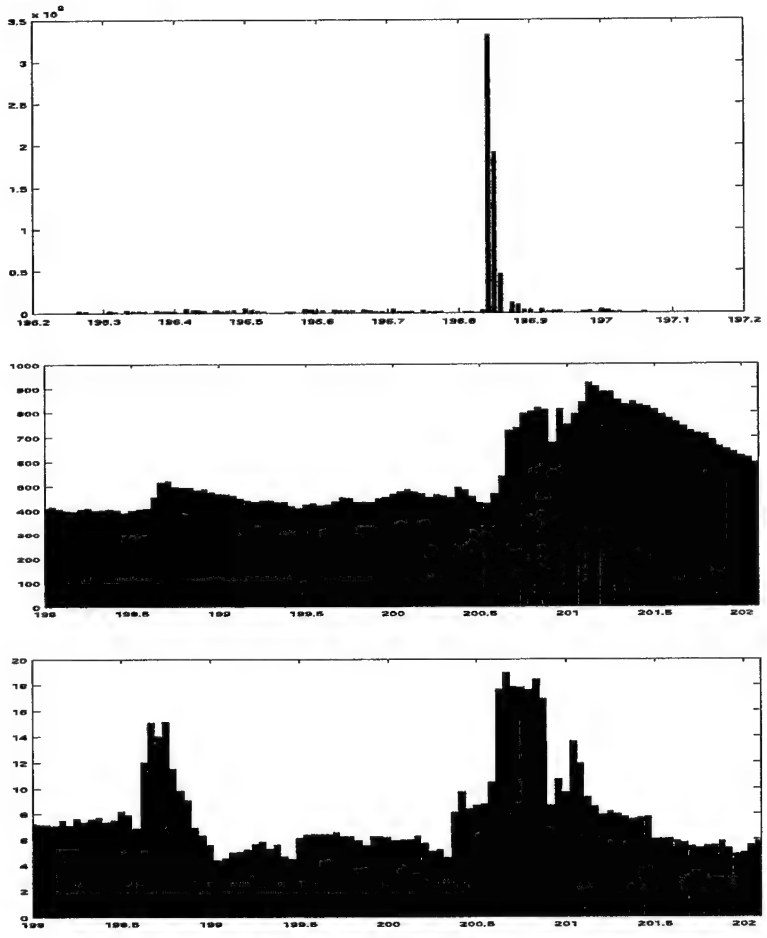


**Figure 12. Difference vector from the July 15, 2002 X3 flare (top). Empty spaces are missing data. ACE measured bulk solar wind speed (middle) and total magnetic field (bottom) associated with this event.**

Using this process of data reduction, we could encode sequences of changes in images occurring over extended periods of time by simply stacking consecutive reduced vectors. The stacking meant that indices stored information about where the block was situated both in the image and in time. The final input to SVM was a set of indices and corresponding values from, on average, 9 consecutive EIT 195 angstrom images, which represented just less than 2 hours of data.

Considering the extent of overlapping surges in ACE data and since directly calculating the travel time of solar wind is a task beyond our immediate capabilities, we felt we couldn't be certain which specific sets of images should be associated with the various jumps in ACE data seen throughout the data set.

To avoid complications of deciphering which of many neighboring fluctuations in data originated at what times, we took a limited definition of an event as being the strongest local increase greater than 20% over 12 hours for bulk solar wind speed or total magnitude of the IMF. We created an automated process that first generated a list of all acceptably large jumps in the ACE measured wind speed and IMF magnitude and then refined the list by extracting only the largest event in the neighborhood. Neighboring events were those which had overlapping images as input.

For this proof-of-concept study, once we established the time of a given event in ACE data we looked at all images available to us within two to six days prior (to account for the uncertain solar wind propagation time) and picked the consecutive pair that changed the most. Pictures from roughly 45 minutes before and after the selected pair were added as a buffer and the entire set was called an event set. We limited our analysis to data from 2002, where we had the best coverage from the EIT standpoint. In all, we identified 28 events for bulk speed increases and 63 for the IMF magnitude jumps. Non-event sets were established as well by looking for periods on the ACE tape where very low wind activity was noted and associating with them sets of consecutive images from the preceding days.

Figure 13 shows examples of bar graphs of the reduced difference vectors for input to SVM along with the IMF magnitude and SWEPAM bulk wind speed data from two to six days following the day of year of the fastest changing image. Data from 15 identified event sets are shown. The input difference vector values are dimensionless scalars. For the graphs of IMF magnitude and wind speed, the x-axis is the day of year for 2002 and y-axes are nano-Tesla and kilometers per second respectively. Figure 14 shows the same parameters for several non-event sets.

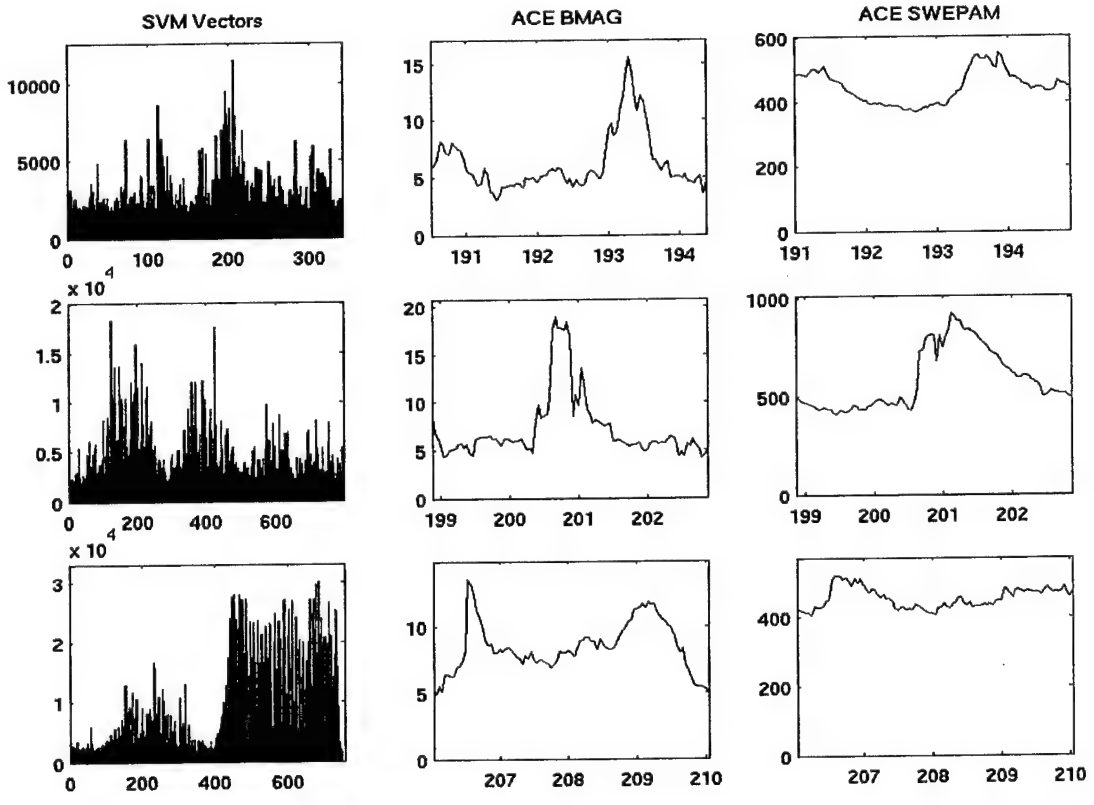



**Figure 13. Examples of reduced difference vectors for input to SVM (left); IMF magnitude (middle); and SWEPAM bulk wind speed (right) from two to six days following the day of year of the fastest changing image.**

**Figure 13 (Cont'd)**

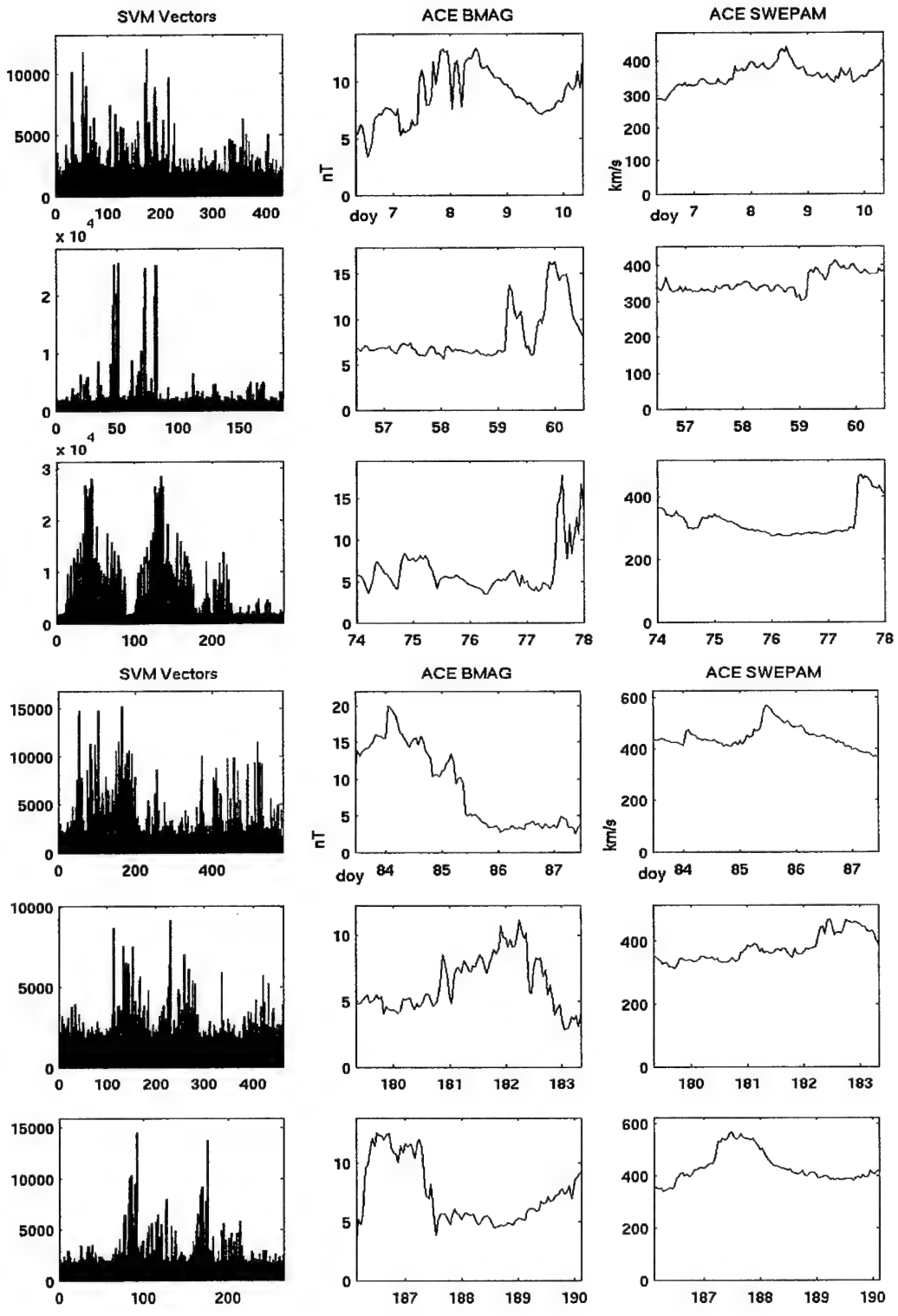
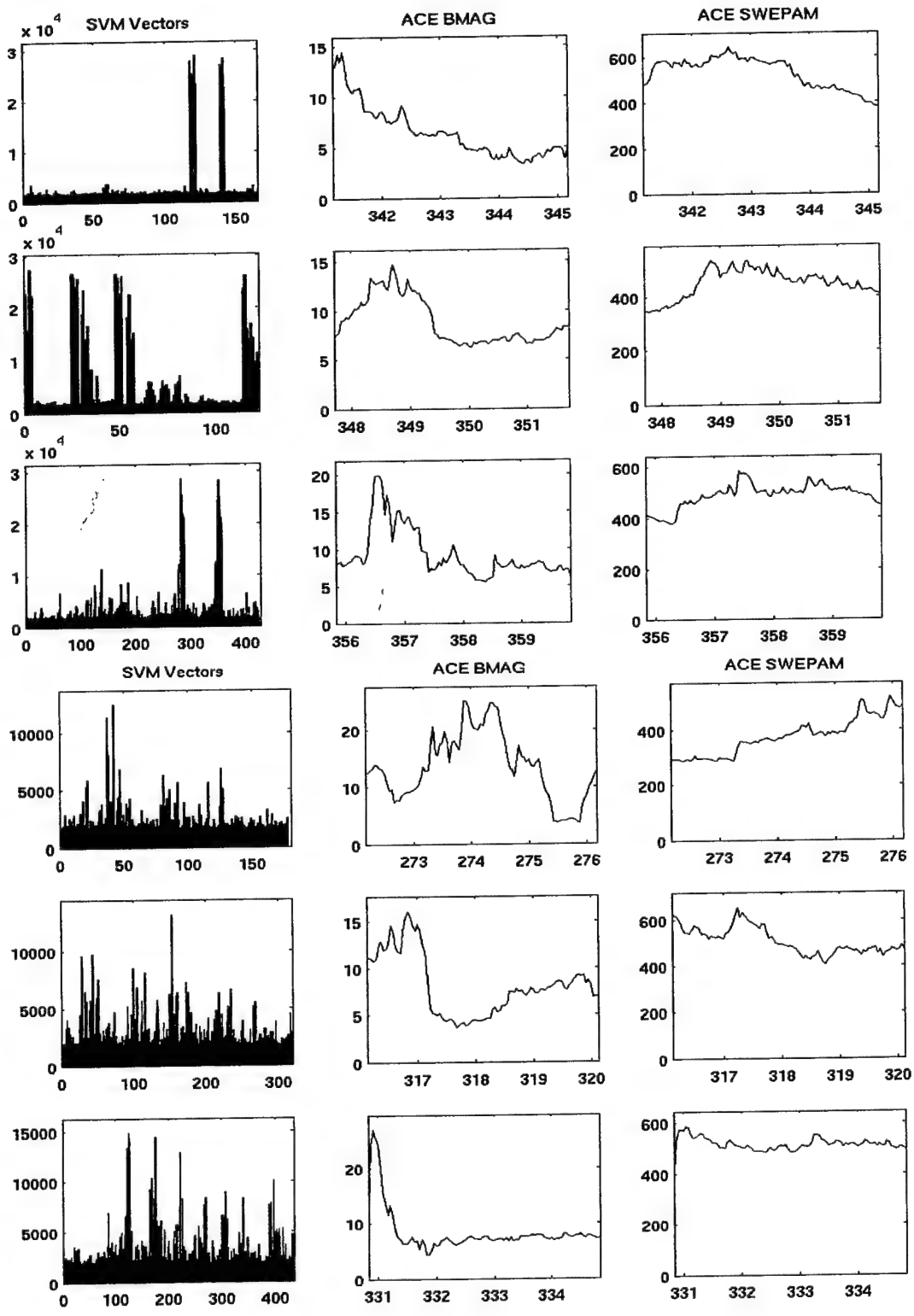


**Figure 13 (Cont'd)**



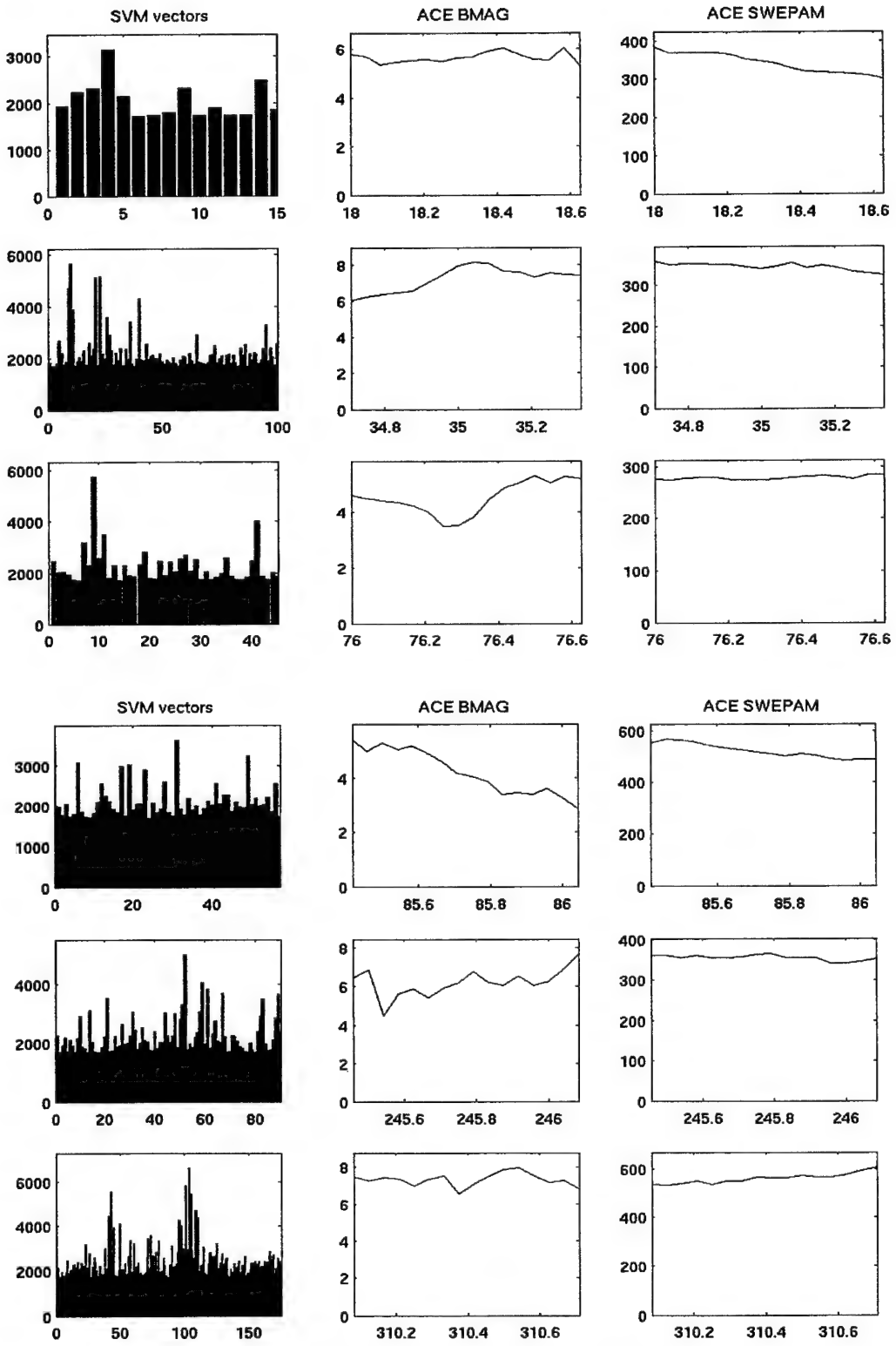



**Figure 14. Similar to Figure 13 but for non-event sets.**

Once the input data has been established for SVM, the task of finding the right kernel function and corresponding set of kernel parameters follows. It is generally accepted that there are no greatly efficient methods of finding the right SVM settings other than through a standard grid search of the parameters of various kernels. In general, there are two parameters to a set: a parameter specific to the given type of kernel and a parameter that determines the acceptable margin of separation between classes. The interaction between these settings is not always intuitive, so it is always suggested to do a broad search through them.

In our case, we found that the polynomial kernel best suited our data. When we did a grid search using various polynomial powers, we found that values between 0.4 and 0.8 for the degree of the kernel did best. In searching the parameters of the current set, we found C, the parameter for setting the margin, was best set to 0.1 and gamma, the polynomial power, was best as 0.58.

In determining the accuracy of the results in SVM training/classification, a method known as cross validation is considered reliable. Cross validation involves separating the full set of data into five subsets. Then each subset alternates as the control prediction set, with the remaining four fifths used for training. Once the process is complete, all sets will have been used for training and all sets will be classified. Moreover, the average of the overall prediction results is a good indicator of the success of future predictions. Below are the results we obtained with the current method of evaluating data.

| | SWEPAM Bulk Wind Speed | IMF Magnitude |
|---|---|---|
| Correctly identified events | 20/28, 71.4% | 42/63, 66.7% |
| False positives | 1/28, 3.5% | 7/63, 11% |

## 3.2 Ensemble Forecasting

We developed a prototype framework for performing ensemble forecasts with the ionospheric model. In order to create the ensembles we chose to perturb two model parameters, F10.7 solar flux and equatorial vertical ExB drift. We then tracked time evolution and spatial distribution of the resulting forecasts and final forecasts uncertainties.

One of the main objectives was to establish "good" behavior of the system. If after reasonable perturbations (~100%) of the model parameters different trajectories showed drastically different results, the ensemble method would be clearly inappropriate to this problem. At the present time the ensemble size is rather small; to keep computations practical we limited the number of ensemble members to less than 10.

Figures 15 shows trajectories of several ensemble members where the equatorial ExB drifts and F10.7 flux were perturbed. As one can see the ensemble members show reasonable spread with absolute values of the spread being larger where the TEC values are larger. Interestingly, however, the relative values of the spread show a different variation effectively pointing to regions of "good" and "bad" predictability. The spread patterns are clearly rather different for these two parameters effectively indicating that a practically useful ensemble forecasting system must include simultaneous variations in both parameters.

Figures 16 and 17 show computed probability density functions for a given location at different local times. The PDF shapes are again very different for the two perturbation choices and, in fact, often show non-Gaussian behavior due to strong non-linearity. The time evolution of the corresponding variance is shown in Figure 18 for both cases. The time behavior of the uncertainties demonstrates sharp changes between relatively "good" and "bad" predictability.

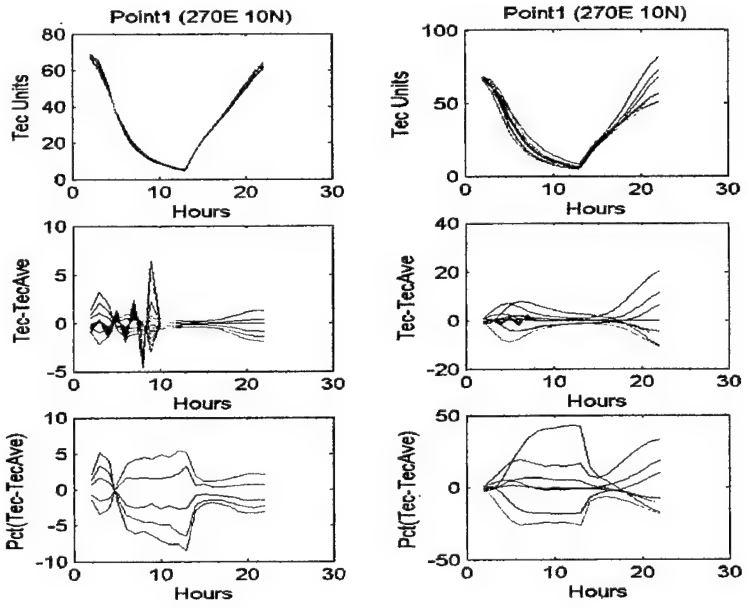



**Figure 15. Ensemble runs with perturbed F10.7 flux (left) and equatorial ExB drifts (right).**

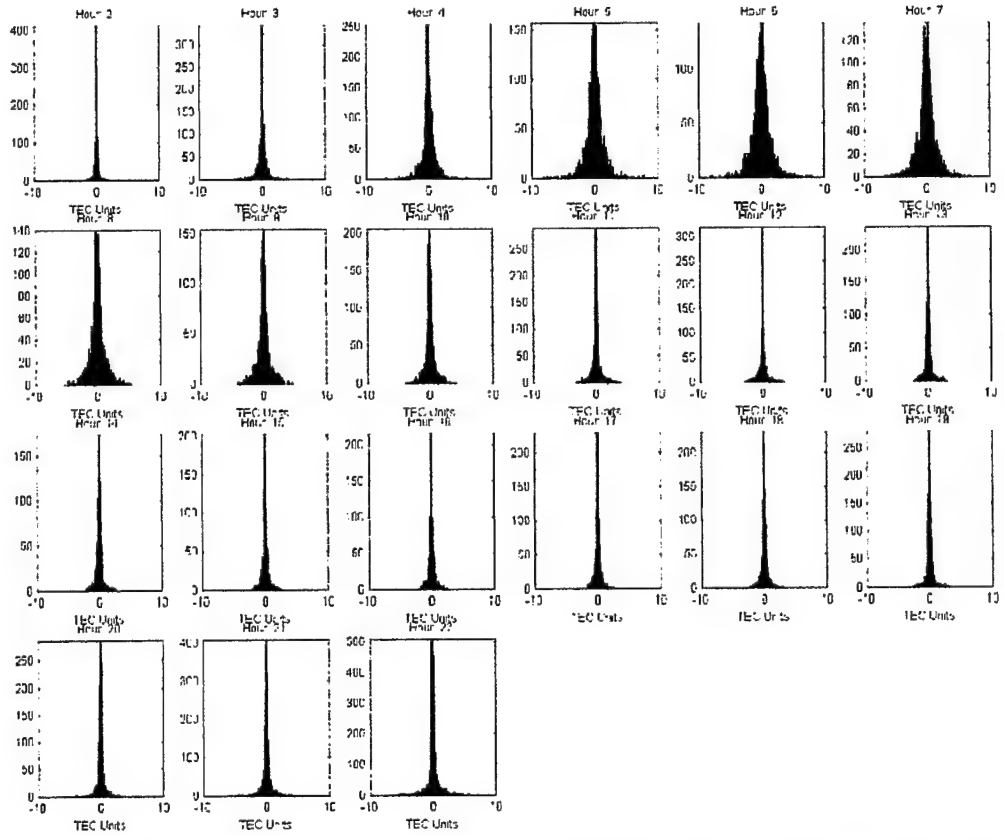



**Figure 16. Time evolution of the total electron content PDF due to perturbed F10.7 flux.**

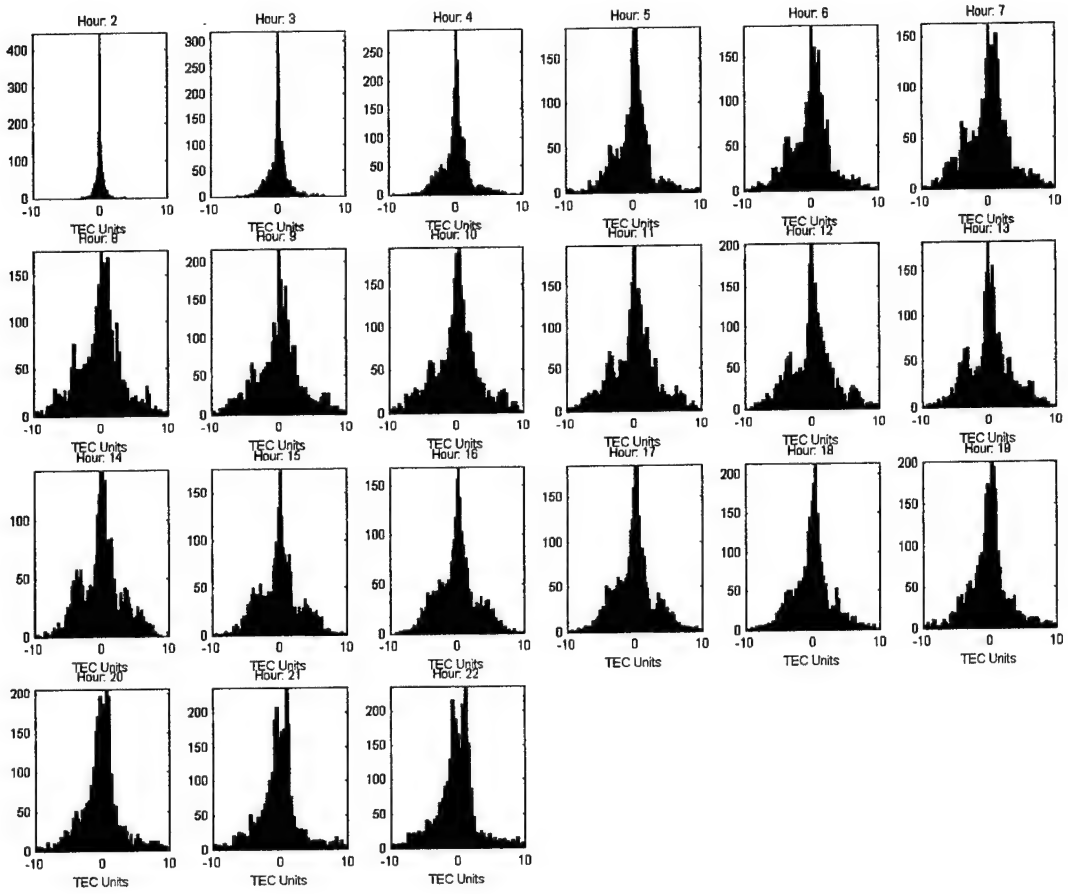



**Figure 17. Time evolution of the total electron content PDF due to perturbed ExB drift.**

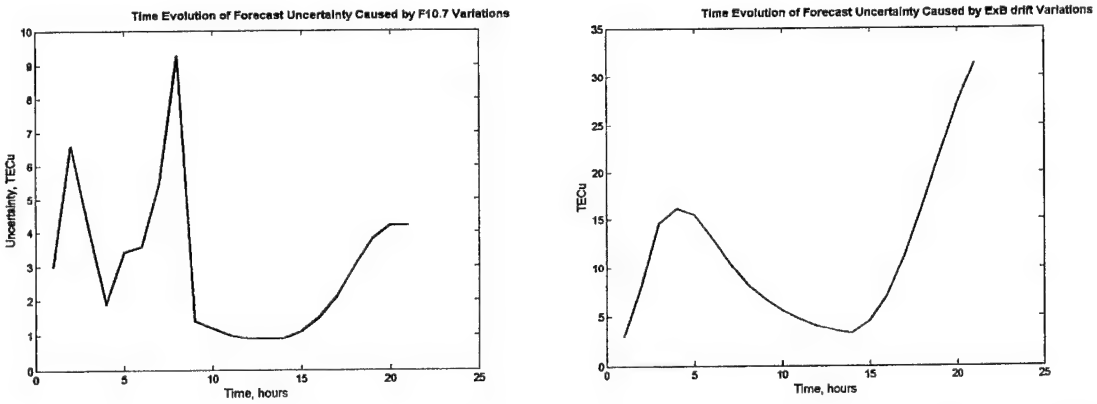



**Figure 18. Time evolution of TEC variance at a particular location due to uncertainty in the F10.7 flux and ExB equatorial drift.**

So far we considered time evolution of forecast PDF at a given location. It is also useful to examine spatial distribution of the PDF at a given moment in time. The next figure demonstrates an example of instantaneous spatial distribution of the spread of the ensemble for the F10.7 ensemble and equatorial drift ensemble for the same time.

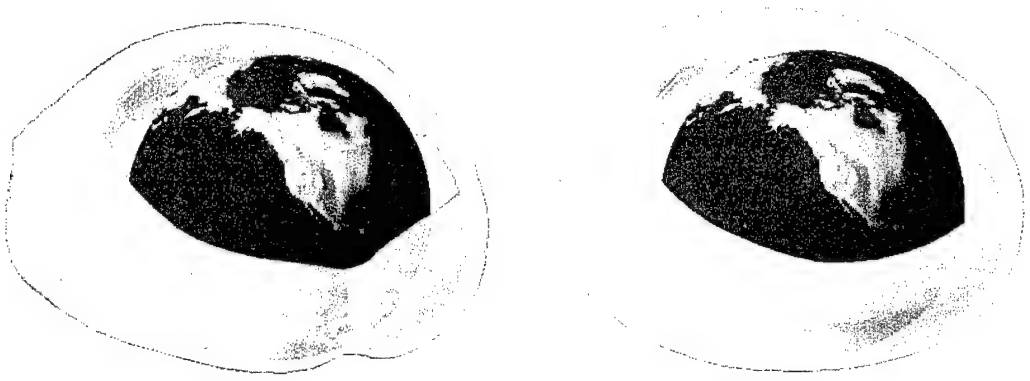


**Figure 19. Instantaneous spatial distribution of ensemble forecasts spread for an ensemble with perturbed solar flux (left) and perturbed equatorial vertical drift (right).**

The forecasts have quite different sensitivity to these two parameters at different geographical locations. As was shown in traditional weather prediction the ensemble forecast quality depends rather dramatically on the number of diverse varying parameters. In the future research we will aim to incorporate as many different variables as our CPU resources would allow.

In summary, we have developed a computational framework and software infrastructure for performing ensemble simulations with perturbed inputs to our ionospheric specification model. We also developed a set of diagnostics for identifying and analyzing characteristics of the forecast probability density function. These diagnostics will be produced in a quasi-operational fashion in the course of the Phase II project.

## 3.3 Operational implementation and integration issues

Without doubt, the greatest obstacle to implementing the forecast system operationally is the latency and poor quality of the SOHO EIT imagery. While in normal operations the images are taken about every 12 minutes, they often do not become available on the EIT ftp site until hours to days later. The delays, as we learned after processing data continuously for several months, vary in an irregular manner. Often the images become available out of order. Since SVM needs a time-ordered sequence of images for classification, the system has to wait until all images are processed. Additionally, quite often many images are missing in a particular sequence. This complicates matters even more as it is impossible to guess whether the system should wait for missing images or analyze the sequence.

A very large portion of time allotted for this project was spent on issues related to obtaining and pre-processing SOHO EIT data. This was something we did not anticipate at the start of the project. In order to obtain a suitably trained SVM, we need to work with as many images as possible. However, when we began, we faced bandwidth and storage limitations and could not download all images from http://umbra.nascom.nasa.gov/eit. To narrow the field, we decided to concentrate on all 195 Å wavelength EIT images of size (1024x1024) and (512x512) for the years 1999-2003. In theory, this roughly corresponds to one image for each 12-minute interval. However, the images are not so neatly distributed in time, especially if one eliminates images with camera errors and missing blocks. Indeed, about 40 percent of all images we obtained were unusable for these reasons.

Indeed, even obtaining a large volume of only 195 Å images was a tedious task. Initially, we ran into FTP time-out errors. Dr. Joseph B. Gurman, a solar scientist at Goddard Space Flight Center, kindly sent us a year's worth of EIT images burned on DVDs by mail. His system administrator helped us with our FTP problem. After that, we were able to use the data Dr. Gurman sent, and write a script to automatically download those images not provided by Dr. Gurman. The current training set is composed of over 200 GB of calibrated SOHO EIT 195 Å images at full resolution spanning several years at about 12 min cadence and the corresponding ACE instrument measurements.

Dr. Gurman also informed us that the images needed to be adjusted to compensate for factors such as CCD aging and stray light. Indeed, our processing of un-calibrated images had been negatively affecting our preliminary results. Fortunately, we installed a program using IDL-based SolarSoft routines written by many NASA researchers to properly calibrate the raw images. Our results subsequently improved significantly as demonstrated by Figure 20.

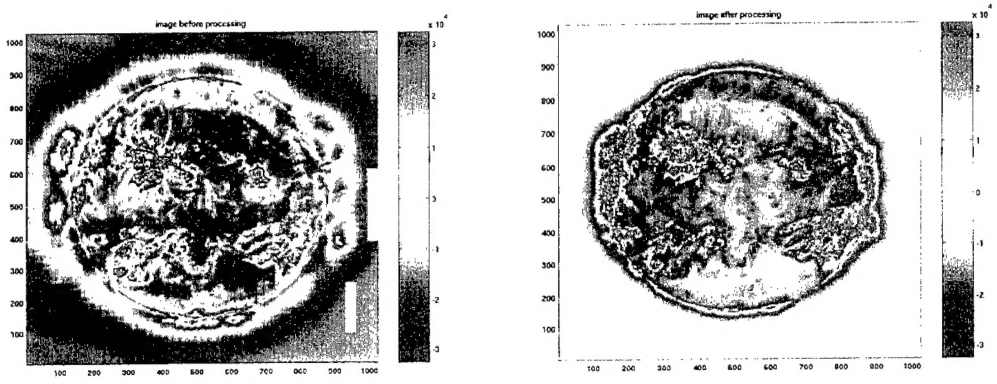



**Figure 20. Un-calibrated (left) and calibrated (right) EIT 195 A images.**

This complicates matters even more, as SVM training and classification is performed by C/C++ code, post processing is done in Matlab, and calibration has to be done in IDL. A number of asynchronous shell and perl scripts "orchestrate" data movement from NASA's EIT ftp server to a local dataset and between these processing components.

## 4. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

The two main conclusions of this feasibility study are that: (1) SVM classifiers show significant promise as a component of a practical automated long-term forecasting system; and (2) ensemble-based PDF evolution scheme for ionospheric system demonstrated significant variability in the forecast spread (uncertainty) in both space and time; the temporal and spatial distributions of related PDFs should benefit the end users if they are operationally computed and promptly distributed.

In the future Phase II work we will extend the SVM classifier development and the ensemble forecasting as described below.

In the classifier training we did not differentiate between events on the periphery of disk and those occurring directly toward Earth. A known use of SVM, called DDAG (Decision Directed Acyclic Graph) SVM, could help improve results as it allows for differentiating between several classes of vectors.

One of the coming steps will involve applying a metric to the difference scheme that would take into account where activity was happening in the image. The metric might have to be different for IMF and bulk wind speed tracking, but since all of our processes are computationally reasonable, tracking with more than one SVM model or method of data extraction will be within our reach. Halo images may also be incorporated to give a more complete the view of an event.

So far we have only used one wavelength of one type of images. Corrupt or missing data are one of the greatest obstacles to a consistent prediction method. Using more wavelengths, most likely with a normalizing metric, could be a way to maintain a level of activity assessment when the various systems came offline from time to time.

We started with 195 Å images because we were advised by scientists at the Air Force Research Laboratory that they were considered the most useful in general. Nonetheless, in the future, we will explore methods for processing images taken at the three other wavelengths: 171Å, 284Å, and 304Å. One of our co-investigators (Dr. Murphy) has developed provably optimal algorithms to test how images at different wavelengths correlate, during his previous employment in an image compression company.

We are still researching the best feature extraction methods. We believe that best methods may come from ideas in Gibson and Low (2000). As illustrated in Figure 21, CMEs are often associated with S-shaped structures. We believe that training the classifier to properly identify them might significantly increase the forecast quality.

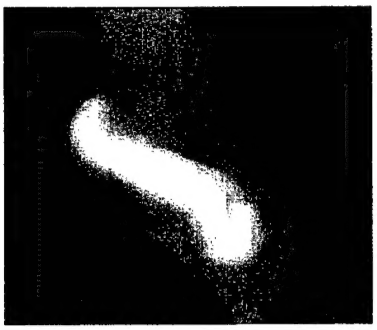


**Figure 21 A typical sigmoidal shape associated with a CME from Gibson and Low (2000).**

However, more research is required to verify this. Once we have that, we would like to optimize our classifiers to detect quiet times as well as intense solar activity periods.

Finally, we would like to extend the classification using related data sources, including solar data in the X-ray region and from ground-based solar observatories.

Use of ensembles of simulations to study ionospheric behavior and to derive temporal and spatial variability of forecast uncertainties is essentially a completely undeveloped field. Given our preliminary results we believe that this work needs to be extended in several directions and that ensemble calculations should be performed systematically. As clearly demonstrated recently in traditional weather prediction, ensemble forecast quality is higher than that of a deterministic forecast and, perhaps more importantly, economic value of ensemble forecast is significantly higher than that of traditional forecasts. The following figures from Zhu et al., Bulletin of American Meteorological Society, January 2002) clearly demonstrate this advantage.

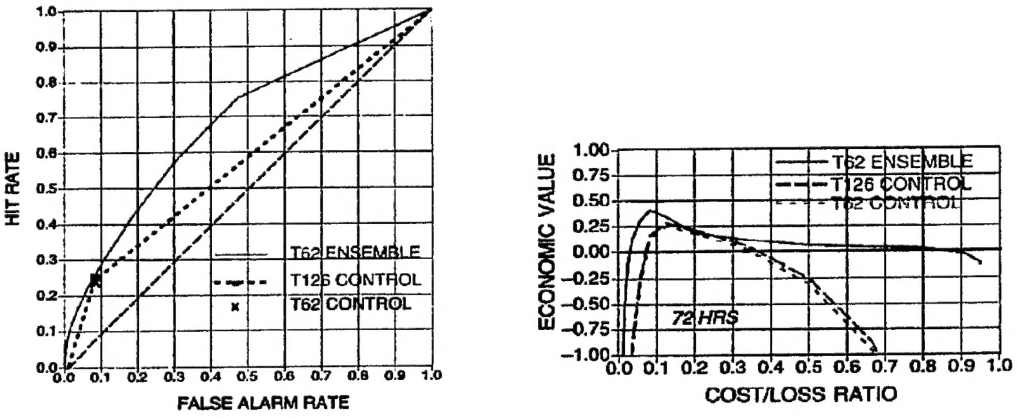



**Figure 22. Comparisons of false alarm rate and economic efficiency of ensemble and deterministic forecasts.**

We plan to take the following steps in the Phase II work:

implement systematic ensemble simulations and forecasting using several different perturbed parameters: solar flux, ExB drift, and model distributions of electron densities;

investigate behavior of the electron density PDF for each of these perturbed runs and for a combination of these simulations;

implement systematic calculations of singular vectors pointing to regions where instabilities growth is the strongest and use this information for developing targeted observational campaigns;

start using ensemble filters for assimilation of TEC data and compare results with extended filter assimilation scheme;

improve our computing infrastructure and increase both the ensemble size and spatial resolution;

## REFERENCES


Canizares, R., 1999: On the application of data assimilation in regional coastal models. Ph.D. thesis, Delft University of Technology, Delft Netherlands, 133 pp.

Cohn, S. E., and R. Todling, 1995: Approximate Kalman filters for unstable dynamics. Second Int. Symp. on Assimilation of Observations in Meteorology and Oceanography, Tokyo, Japan, WMO, 241–246.

Cohn, S. E., and R. Todling, 1996: Approximate data assimilation schemes for stable and unstable dynamics. J. Meteor. Soc. Japan, 74, 63–75.

Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S. H. (1999). Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification, Proteins 35 , 401-7.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99** (C5), 10 143–10 162.

Evensen, G., and P. J. van Leeuwen, 1996: Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.*, **124**, 85–96.

Fuller-Rowell, T. J., D. Rees, S. Quegan, R. J. Moffett, and G. J. Bailey, Interaction between neutral thermospheric composition and the polar ionosphere using a coupled ionosphere-thermosphere model. *J. Geophys. Res.*, **92**, 7744, 1987.

Houtekamer, P. L., and A. L. Mitchel, 1998: Data assimilation using ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811.

Jaakkola, T., Diekhans, M., and Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies, Proc Int Conf Intell Syst Mol Biol, 149-58.

Lermusiaux, P. F. J., 1997: Error subspace data assimilation methods for ocean field estimation: Theory, validation and applications. Ph. D. thesis, Harvard University, 402 pp.

Matsuo, T., A. D. Richmond, and D. W. Nychka, Modes of high-latitude electric field variability derived from DE-2 measurements: Empirical Orthogonal Function (EOF) analysis, *Geophys. Res. Lett.*, **29**(7), 2002.

Mukherjee, S., E. Osuna and F. Girosi, Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines, In Proc. of IEEE NNSP'97, Amelia Island, FL, September 1997, pp. 511-519.

Osuna, J., Soberon, X., and Morett, E. (1997). A proposed architecture for the central domain of the bacterial enhancer-binding proteins based on secondary structure prediction and fold recognition, Protein Sci 6 , 543-55.

Osuna. E., R. Freund and F. Girosi, Training Support Vector Machines: an Application to Face Detection, Proceedings of CVPR'97, June 17-19,1997, Puerto Rico.

Pham, D., J. Verron, and M. Rouband, 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Syst.*, **16**, 323–340.

Richmond, A. D., and Y. Kamide, Mapping electrodynamic features of the high-latitude ionosphere from localized observations: Technique, *J. Geophys. Res.*, **93**, 5741-5759, 1988.

Vapnik, V, Statistical Learning Theory. Wiley, New York., 1998.

Verlaan, M., and A. W. Heemink, 1995: Data assimilation schemes for non-linear shallow water flow models. Proc. Second Int. Symp. on Assimilation of Observations, Tokyo, Japan, WMO, 247–252.

Verlaan, M., and A. W. Heemink, 1997: Tidal flow forecasting using reduced-rank square root filters. Stochastic Hydro. Hydraul., 11, 349–368.

Weimer, D. R., An improved model of ionospheric electric potentials including substorm perturbations and application to the Geospace Environment Modeling November 24, 1996, event, *J. Geophys. Res.*, **106**, 407-416, 2001.

Zavaljevski, N., Stevens, F. J., and Reifman, J. (2002). Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions, Bioinformatics 18 , 689-96.

Millward et al., A coupled thermosphere-ionosphere-plasmosphere model (CTIP), in STEP: Handbook of Ionospheric Models, STEP Report, editor R.W. Schunk , 1996.

Bailey and Balan, A low-latitude ionosphere-plasmosphere model, in STEP: Handbook of Ionospheric Models, STEP Report, editor R.W. Schunk , 1996.

Huba et al., SAMI2 is another model of the ionosphere: a new low-latitude ionosphere model, J. *Geophys. Res.*, **23**,035, 2000.

Fuller-Rowell T.J., D. Rees, S. Quegan, R.J. Moffett, M.V. Codrescu, and G.H. Millward, A coupled thermosphere ionosphere model (CTIM). Handbook of Ionospheric Models, STEP Report, editor R.W. Schunk , 1996.

Khattatov, B. V., J. C. Gille, L. V. Lyjak, G. P. Brasseur, V. L. Dvortsov, A. E. Roche, and J. Waters, Assimilation of photochemically active species and a case analysis of UARS data., *J. Geophys. Res.*, **104**, 18,715-18,737, 1999.